

# On norm sub-additivity and super-additivity inequalities for concave and convex functions

Koenraad M.R. Audenaert

*Department of Mathematics, Royal Holloway, University of London,  
Egham TW20 0EX, United Kingdom*

Jaspal Singh Aujla

*Department of Mathematics, National Institute of Technology,  
Jalandhar 144011, Punjab, India*

---

## Abstract

Sub-additive and super-additive inequalities for concave and convex functions have been generalized to the case of matrices by several authors over a period of time. These lead to some interesting inequalities for matrices, which in some cases coincide with, and in other cases are at variance with the corresponding inequalities for real numbers. We survey some of these matrix inequalities and do further investigations into these.

We introduce the novel notion of dominated majorization between the spectra of two Hermitian matrices  $B$  and  $C$ , dominated by a third Hermitian matrix  $A$ . Based on an explicit formula for the gradient of the sum of the  $k$  largest eigenvalues of a Hermitian matrix, we show that under certain conditions dominated majorization reduces to a linear majorization-like relation between the diagonal elements of  $B$  and  $C$  in a certain basis. We use this notion as a tool to give new, elementary proofs for the sub-additivity inequality for non-negative concave functions first proved by Bourin and Uchiyama and the corresponding super-additivity inequality for non-negative convex functions first proven by Kosem.

Finally, we present counterexamples to some conjectures that Ando's inequality for operator convex functions could more generally hold, e.g. for ordinary convex, non-negative functions.

*Dedicated to the memory of Ky Fan*

*Key words:* Matrix Norm Inequality, Positive Semidefinite Matrix, Convex function, Majorization

*1991 MSC:* 15A60

---

## 1 Introduction

Two of the basic properties that a real-valued function  $f(x)$  defined over the reals can possess are sub-additivity and super-additivity. Sub-additivity means that for all  $x, y$  in the domain of  $f$ ,

$$f(x + y) \leq f(x) + f(y),$$

while super-additivity means the opposite

$$f(x) + f(y) \leq f(x + y).$$

Two classical theorems that characterise sub- and super-additivity for functions defined on  $\mathbb{R}^+$  (although not completely) are presented as Theorem 7.2.4 and 7.2.5 in [12]. Their Theorem 7.2.4 states that functions  $f$  for which  $f(t)/t$  is decreasing in  $\mathbb{R}_+$  are subadditive. Theorem 7.2.5 in [12] states that any measurable concave function  $f$  is subadditive in  $\mathbb{R}_+$  iff  $f(0+) \geq 0$ .

In recent years, ongoing effort has been spent to characterise matrix functions exhibiting similar sub-additivity or super-additivity properties. Of course, many variations on this theme are possible, and in this paper we restrict attention to sub- and super-additivity in norm for non-negative functions. For a given unitarily invariant norm  $||| \cdot |||$ , these amount to the norm inequalities  $|||f(A) + f(B)||| \leq |||f(A + B)|||$  (or reversed), with positive semidefinite  $A$  and  $B$ , but one can equally well consider the inequality  $|||f(A) - f(B)||| \leq |||f(|A - B|)|||$  (or reversed). Historically, these inequalities have been proven first for operator monotone, and/or operator concave functions  $f$ , and only later have they been generalised to non-negative functions that are concave and/or convex. Interestingly, the proofs of these generalisations exploit the corresponding results for operator monotone/concave functions.

In this paper we first give a historical overview of these developments, in Sections 3 and 4. Then we resolve a number of still open questions regarding the inequality  $|||f(|A - B|)||| \leq |||f(A) - f(B)|||$ , which is known to be true for operator convex functions. We show by counterexample that it does not hold in general for non-negative convex functions, nor do a number of successively weakened versions. By imposing the condition  $A \geq \|B\|_\infty$ , we obtain the closest match of this inequality that does hold for convex functions (or in reversed sense for concave functions), namely the eigenvalue inequality  $\lambda_k^\downarrow(f(A - B)) \leq \lambda_k^\downarrow(f(A) - f(B))$ , for all  $k$ .

In Section 6, we present a new and elementary proof of a sub-additivity norm inequality for non-negative concave functions and a super-additivity norm in-

---

*Email addresses:* `koenraad.audenaert@rhul.ac.uk` (Koenraad M.R. Audenaert), `aujla@nitj.ac.in` (Jaspal Singh Aujla).

equality for non-negative convex functions, that do not rely on the corresponding inequality for operator monotone/convex functions, nor on the theory of operator monotone functions. The proof exploits the novel notion of dominated majorization between the spectra of two Hermitian matrices  $B$  and  $C$ , dominated by a third Hermitian matrix  $A$ . Based on an explicit formula for the gradient of the sum of the  $k$  largest eigenvalues of a Hermitian matrix, we show that under certain conditions this dominated majorization reduces to a linear majorization-like relation between the diagonal elements of  $B$  and  $C$  in a certain basis. This is explained in full detail in Section 5 (with one of the proofs postponed to Section 7).

## 2 Preliminaries

In this section, we introduce the notations and necessary prerequisites; a more detailed exposition can be found, e.g. in [7].

Throughout,  $\mathbb{M}_n$  shall denote the set of  $n \times n$  complex matrices and  $\mathbb{M}_n^H$  shall denote the set of all Hermitian matrices in  $\mathbb{M}_n$ . We shall abbreviate the terms positive semidefinite and positive definite by PSD and PD, respectively. By  $A \geq B$ , we mean that  $A - B \geq 0$ . Let  $I$  be an interval in  $\mathbb{R}$ . We shall denote by  $\mathbb{M}_n^H(I)$ , the set of all Hermitian matrices in  $\mathbb{M}_n$  whose spectrum is contained in the interval  $I$ .

We denote the identity matrix by  $\mathbf{I}$ , and use the shorthand  $a = a\mathbf{I}$  for scalar matrices.

We denote the absolute value by  $|\cdot|$ , both for scalars and for matrices. For matrices this is defined as  $|A| := (A^*A)^{1/2}$ . Similarly, we denote the positive part of a real scalar or Hermitian matrix by  $(\cdot)_+$ , and define it by  $A_+ := (A + |A|)/2$ . We denote the vector of diagonal entries of a matrix  $A$  by  $\text{Diag}(A)$ . We will use the abbreviations LHS and RHS for left-hand side and right-hand side, respectively.

Let  $A \in \mathbb{M}_n^H(I)$  have the spectral decomposition

$$A = U^* \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) U$$

where  $U$  is a unitary matrix and  $\lambda_1, \lambda_2, \dots, \lambda_n$  are the eigenvalues of  $A$ . Let  $f$  be a real valued function defined on  $I$ . Then  $f(A)$  is defined by

$$f(A) = U^* \text{diag}(f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n)) U.$$

Let  $n \in \mathbb{N}$  be arbitrary but fixed. The function  $f$  is called *matrix monotone*

of order  $n$  on  $I$  if

$$A \geq B \implies f(A) \geq f(B)$$

for all  $A, B \in \mathbb{M}_n^H(I)$ , and *matrix convex of order  $n$  on  $I$*  if

$$f(\alpha A + (1 - \alpha)B) \leq \alpha f(A) + (1 - \alpha)f(B)$$

for all  $0 \leq \alpha \leq 1$  and  $A, B \in \mathbb{M}_n^H(I)$ . Likewise,  $f$  is called *matrix concave of order  $n$  on  $I$*  if  $-f$  is matrix convex of order  $n$  on  $I$ . If the function  $f$  is matrix monotone of all orders  $n$  on  $I$  then  $f$  is called *operator monotone on  $I$* . The operator convexity and operator concavity are defined similarly.

A norm  $||| \cdot |||$  on  $\mathbb{M}_n$  is called unitarily invariant (UI) or symmetric if

$$|||UAV||| = |||A|||$$

for all  $A \in \mathbb{M}_n$  and for all unitary  $U, V \in \mathbb{M}_n$ . The most basic unitarily invariant norms are the Ky Fan norms  $|| \cdot ||_{(k)}$ ,  $(k = 1, 2, \dots, n)$ , defined as

$$||A||_{(k)} = \sum_{j=1}^k \sigma_j(A), \quad (k = 1, 2, \dots, n)$$

and the Schatten  $p$ -norms defined as

$$||A||_p = \left( \sum_{j=1}^n (\sigma_j(A))^p \right)^{1/p},$$

$1 \leq p < \infty$ , where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  are the singular values of  $A \in \mathbb{M}_n$ , that is, the eigenvalues of  $|A|$ . The spectral norm (or operator norm) is given by  $||A||_\infty = s_1(A) = \lim_{p \rightarrow \infty} ||A||_p$ .

The famous Ky Fan dominance theorem states that a matrix  $B$  dominates another matrix  $A$  in all UI norms if and only if it does so in all Ky Fan norms. The latter set of relations can be written as a weak majorization relation between the vectors of singular values of  $A$  and  $B$ :

$$\sigma^\downarrow(A) \prec_w \sigma^\downarrow(B) : \quad \sum_{j=1}^k \sigma_j(A) \leq \sum_{j=1}^k \sigma_j(B), \quad 1 \leq k \leq n.$$

For PSD matrices, the above domination relation translates to a weak majorization between the vectors of eigenvalues:  $\lambda^\downarrow(A) \prec_w \lambda^\downarrow(B)$ . Here,  $\lambda^\downarrow(A)$  denotes the (real) vector of eigenvalues of  $A$  sorted in non-increasing order.

Weyl's monotonicity theorem ([7], Corollary III.2.3) states that

$$\lambda_k^\downarrow(A) \leq \lambda_k^\downarrow(A + B), \quad 1 \leq k \leq n,$$

for Hermitian  $A$  and PSD  $B$ .

Finally, we refer the reader to Chapter 2 of [14] for an exposition of a number of important functional analytic properties of eigenvalues and corresponding eigenspaces of a Hermitian matrix, which we will need in the proof of Theorem 2.

### 3 Comparison of norms $|||f(A) + f(B)|||$ and $|||f(A + B)|||$

For PD matrices  $A, B$ , McCarthy [19] proved that

$$\|A^r + B^r\|_1 \leq \|(A + B)^r\|_1, \quad 1 \leq r < \infty$$

and

$$\|A^r + B^r\|_1 \geq \|(A + B)^r\|_1, \quad 0 \leq r \leq 1.$$

Bhatia and Kittaneh [8] proved the above-mentioned inequalities for the operator norm. There they also proved that

$$|||A^m + B^m||| \leq |||(A + B)^m|||, \quad m = 1, 2, \dots$$

for  $A, B \geq 0$  and conjectured that if  $f$  is operator monotone function on  $[0, \infty)$  with  $f(0) = 0$  then

$$|||f(A + B)||| \leq |||f(A) + f(B)|||. \quad (1)$$

Hiai also posed this conjecture in [11]. Ando and Zhan affirmatively settled this conjecture in [2]. As a corollary they obtained that if  $f$  is an increasing function on  $[0, \infty)$  with  $f(0) = 0$ ,  $f(\infty) = \infty$  and if the inverse function of  $f$  is operator monotone then

$$|||f(A + B)||| \geq |||f(A) + f(B)|||. \quad (2)$$

Since the inverse function of a non-negative operator convex function on  $[0, \infty)$  with  $f(0) = 0$  is operator monotone [1], we conclude that inequality (2) holds for any operator convex function on  $[0, \infty)$  with  $f(0) = 0$ . In [5] it was shown that if the non-negative functions  $f, g$  on  $[0, \infty)$  satisfy inequality (2) then the functions  $f + g$ ,  $f \circ g$  and  $fg$  also satisfy (2). It was further shown that any polynomial  $p$  with non-negative coefficients and  $p(0) = 0$  satisfy (2).

This prompted the authors to conjecture in [5] that any non-negative convex function on  $[0, \infty)$  with  $f(0) = 0$  should also satisfy (2). Note that such functions must automatically be increasing functions. Using the fact that a non-negative convex function on  $[0, \infty)$  with  $f(0) = 0$  can be approximated uniformly on a finite interval by a positive linear combination of angle functions, Kosem settled this conjecture affirmatively in [17]. Later on Bourin and

Uchiyama proved ([10]; see also [4]) that any non-negative concave function on  $[0, \infty)$  (again such functions must be increasing) satisfies (1).

It is shown in [3,5] that if a non-negative function  $f$  satisfies (1) then it is concave and if it satisfies (2) then it is convex with  $f(0) = 0$ . Hence within the set of non-negative  $f$  these results give a full characterisation of all possible  $f$  satisfying these inequalities. This completes our discussion in this section.

#### 4 Comparison of norms $||f(A) - f(B)||$ and $||f(|A - B|)||$

We begin this section with the inequality of Powers and Størmer [21], derived in the course of their work on free states of the canonical anti-commutation relations. They proved that if  $A, B$  are PSD then

$$||A^{1/2} - B^{1/2}||_2^2 \leq ||A - B||_1.$$

Kittaneh [15] generalized this to show that

$$||A^{1/2} - B^{1/2}||_{2p}^2 \leq ||A - B||_p$$

for  $1 \leq p \leq \infty$ . Note that for any matrix  $T$ , we have  $||T||_{2p}^2 = ||T^*T||_p$ ,  $1 \leq p \leq \infty$ , so this result of Kittaneh can be restated as

$$||(A - B)^2||_p \leq ||A^2 - B^2||_p.$$

Bhatia [6] proved this inequality for all unitarily invariant norms. There he also proved that

$$||(A - B)^{2^k}||_p \leq ||A^{2^k} - B^{2^k}||_p, \quad k = 1, 2, \dots$$

The above inequality when specialized to the  $p$ -norms gives

$$||A^{1/m} - B^{1/m}||_{mp}^m \leq ||A - B||_p$$

for all integers  $m$  of the form  $2^k$ ,  $k = 1, 2, \dots$ , which is an interesting generalisation of the Powers-Størmer inequality.

In [9] Birman, Koplienko and Solomyak proved that

$$||A^r - B^r||_\infty \leq |||A - B|^r||_\infty, \quad 0 \leq r \leq 1$$

for all  $A, B \geq 0$ . Note that the function  $f(x) = x^r$  is operator monotone on  $[0, \infty)$ . This motivated Kittaneh and Kosaki [16] to prove that if  $f$  is non-negative operator monotone on  $[0, \infty)$  then

$$||f(A) - f(B)||_\infty \leq f(||A - B||_\infty) = ||f(|A - B|)||_\infty$$

for all  $A, B \geq 0$ . Then Ando [1] proved that if  $f$  is non-negative operator monotone on  $[0, \infty)$  then

$$|||f(A) - f(B)||| \leq |||f(|A - B|)||| \quad (3)$$

$A, B \geq 0$ , for all unitarily invariant norms. As a corollary to this result, Ando deduced that the reverse inequality holds for all functions  $f$  on  $[0, \infty)$  with  $f(0) = 0$  and  $f(\infty) = \infty$  if the inverse function of  $f$  is operator monotone. Since the inverse function of a non-negative operator convex function on  $[0, \infty)$  with  $f(0) = 0$  is operator monotone [1], we conclude that if  $f$  is operator convex on  $[0, \infty)$  with  $f(0) = 0$  then we have

$$|||f(A) - f(B)||| \geq |||f(|A - B|)|||. \quad (4)$$

Afterward, Mathias [18] proved that the inequality (3) holds for any non-negative matrix monotone function of order  $n$  on  $[0, \infty)$ . One may wonder whether, in a similar vein, inequality (4) can be proved for a non-negative increasing matrix convex function  $f$  of order  $n$  on  $[0, \infty)$  with  $f(0) = 0$ .

We have seen that inequality (1) holds for non-negative increasing concave functions on  $[0, \infty)$  and inequality (2) holds for non-negative increasing convex functions on  $[0, \infty)$  with  $f(0) = 0$ . In the same spirit, we consider the question whether inequalities (3) and (4) can also be generalized to non-negative concave and convex functions respectively. We raise and answer several questions in this direction.

**Question 1** *For all  $A, B \geq 0$ , for all UI norms, and for non-negative increasing convex functions  $g$  on  $[0, \infty)$  with  $g(0) = 0$ , does the inequality  $|||g(A) - g(B)||| \geq |||g(|A - B|)|||$  hold?*

The answer to this question is negative, as shown by the following counterexample. We consider the convex angle function  $g(x) = x + (x - 1)_+$  and the operator norm. For the  $2 \times 2$  PSD matrices

$$A = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.6 \end{pmatrix}, \quad B = \begin{pmatrix} 0.8 & 0.5 \\ 0.5 & 0.4 \end{pmatrix},$$

the eigenvalues of  $g(|A - B|)$  are 0.65249 and 0.35249, while those of  $g(A) - g(B)$  are 0.65010 and  $-0.48862$ . Thus,  $||g(|A - B|)||_\infty = 0.65249$ , which is larger than  $||g(A) - g(B)||_\infty = 0.65010$ .  $\square$

Under the additional restriction  $A \geq B$ , the absolute value in the argument of  $g$  in the RHS vanishes, leading to a simplified statement and a second question, with better hopes for success. Introducing the matrix  $\Delta = A - B$ ,

**Question 2** For all  $B, \Delta \geq 0$ , for all UI norms, and for non-negative increasing convex functions  $g$  on  $[0, \infty)$  with  $g(0) = 0$ , does the inequality  $|||g(B + \Delta) - g(B)||| \geq |||g(\Delta)|||$  hold?

This restricted case also turns out to have a negative answer. Counterexamples, however, were much harder to find, and required a reduction of the problem based on certain results about a novel majorization-like relation, which we call dominated majorization. This will be the subject of Sections 5 and 6, where a number of results of independent interest are proven.

It is also very reasonable to ask:

**Question 3** For all  $B, \Delta \geq 0$ , for all UI norms, and for non-negative increasing concave functions  $f$  on  $[0, \infty)$ , does the inequality  $|||f(B + \Delta) - f(B)||| \leq |||f(\Delta)|||$  hold?

Again, this statement is false, as the following counterexample shows. Consider the concave angle function  $f(x) = \min(x, 1) = x - (x - 1)_+$ , and the  $3 \times 3$  PSD matrices

$$B = \begin{pmatrix} 0.701816 & 0.317887 & 0.198910 \\ 0.317887 & 1.014950 & -0.093826 \\ 0.198910 & -0.093826 & 0.274236 \end{pmatrix}$$

and

$$\Delta = \begin{pmatrix} 0.192713 & 0 & 0 \\ 0 & 0.446505 & 0 \\ 0 & 0 & 0.455416 \end{pmatrix}.$$

One gets

$$||f(\Delta)||_\infty = 0.455416$$

while

$$||f(B + \Delta) - f(B)||_\infty = 0.455776.$$

□

Next we consider an even more restricted special case, in which the inequalities (3) and (4) finally do hold. We actually prove that a stronger relationship holds in this special case. We shall use the notation  $\lambda^\downarrow(X) \leq \lambda^\downarrow(Y)$  whenever  $\lambda_k^\downarrow(X) \leq \lambda_k^\downarrow(Y)$  holds for all  $k$ .

**Theorem 1** For a non-negative, increasing concave function  $g$  on  $[0, \infty)$ , and matrices  $A, B \geq 0$  such that  $A \geq ||B||_\infty$ , we have

$$\lambda^\downarrow(g(A - B)) \geq \lambda^\downarrow(g(A) - g(B)). \quad (5)$$



An easy corollary is the corresponding statement for non-negative convex functions.

**Corollary 1** *Let  $f$  be a non-negative strictly increasing convex function on  $[0, \infty)$  with  $f(0) = 0$ . Let  $A, B \geq 0$  be such that  $A \geq \|B\|_\infty$ . Then*

$$\lambda^\downarrow(f(A - B)) \leq \lambda^\downarrow(f(A) - f(B)). \quad (6)$$

*Proof.* Let  $f = g^{-1}$ , with  $g$  satisfying the conditions of Theorem 1. Upon replacing  $A$  by  $f(A)$  and  $B$  by  $f(B)$ , the condition  $A \geq \|B\|_\infty$  is unharmed as  $f$  is monotonous. Furthermore, (5) becomes

$$\lambda^\downarrow(g(f(A) - f(B))) \geq \lambda^\downarrow(A - B).$$

Applying the function  $f$  on both sides does not change the ordering, again because of monotonicity of  $f$ , and yields validity of inequality (6).  $\square$

These two results obviously imply the corresponding majorization relations, and by Ky Fan dominance, relations in any UI norm.

*Proof of Theorem 1.* W.l.o.g. we will assume  $\|B\|_\infty = 1$ , since any other value can be absorbed in the definition of  $g$ .

It is immediately clear that if (5) holds for  $g$  that in addition satisfy  $g(0) = 0$ , then it must also hold without that constraint, i.e. for functions  $g(x) + c$ , with  $c \geq 0$ . This is because the additional constant  $c$  cancels out in the LHS, while  $\lambda^\downarrow(g(A - B) + c) \geq \lambda^\downarrow(g(A - B))$ .

Furthermore, (5) remains valid when replacing  $g(x)$  with  $ag(x)$ , for  $a > 0$ . Thus, w.l.o.g. we can assume  $g(0) = 0$  and  $g(1) = 1$ . Together with concavity of  $g$ , this implies that, for  $0 \leq x \leq 1$ ,  $g(x) \geq x$ , while for  $x \geq 1$ , the one-sided derivative  $g'(x) \leq 1$  (since concave functions need not be differentiable everywhere, we have to use the one-sided derivative  $g'(x) = \lim_{t \rightarrow 0^+} (g(x+t) - g(x))/t$ ).

Since  $0 \leq B \leq \mathbf{I}$ , and for  $0 \leq x \leq 1$ ,  $g(x) \geq x$  holds, we have  $g(B) \geq B$ , or  $-g(B) \leq -B$ . By Weyl monotonicity, this implies  $\lambda^\downarrow(g(A) - g(B)) \leq \lambda^\downarrow(g(A) - B)$ . Thus, statement (5) would be implied by the stronger statement

$$\lambda^\downarrow(g(A) - B) \leq \lambda^\downarrow(g(A - B)). \quad (7)$$

Now note that the argument of  $g$  in the LHS satisfies  $A \geq \mathbf{I}$ . Thus, in principle,

we could replace  $g(x)$  in the LHS by another function  $h(x)$  defined as

$$h(x) = \begin{cases} g(x), & \text{if } x \geq 1 \\ x, & \text{otherwise.} \end{cases} \quad (8)$$

If we also do that in the RHS, we get a stronger statement than (7). Indeed,  $h(x) \leq g(x)$  for  $x \geq 0$  and  $A - B \geq 0$ , and therefore  $h(A - B) \leq g(A - B)$  holds. By Weyl monotonicity again, we see that (7) is implied by

$$\lambda^\downarrow(h(A) - B) \leq \lambda^\downarrow(h(A - B)). \quad (9)$$

The importance of this move is that  $h(x)$  is still an increasing and concave function (because  $g'(x) \leq 1$  for  $x \geq 1$ ), but now has  $h'(x) \leq 1$  for  $x \geq 0$ .

Defining  $C = A - B$ , which is positive semi-definite, we now have to show the inequality

$$\lambda_k^\downarrow(h(C + B) - B) \leq \lambda_k^\downarrow(h(C)) = h(\lambda_k^\downarrow(C)),$$

for every  $k$ . Fixing  $k$ , and introducing the shorthand  $x_0 = \lambda_k^\downarrow(C)$ , we can exploit concavity of  $h$  to bound it from above as  $h(x) \leq a(x - x_0) + h(x_0)$ , where  $a = h'(x_0) \leq 1$ . Again by Weyl monotonicity, we find

$$\begin{aligned} \lambda_k^\downarrow(h(C + B) - B) &\leq \lambda_k^\downarrow(a(C + B - x_0) + h(x_0) - B) \\ &= \lambda_k^\downarrow(aC + (a - 1)B - ax_0 + h(x_0)) \\ &\leq \lambda_k^\downarrow(aC) - ax_0 + h(x_0) = h(x_0), \end{aligned}$$

where in the second line we could remove the term  $(a - 1)B$  because it is negative. This being true for all  $k$ , we have proved (9) and all previous statements that follow from it, including the statement of the theorem.  $\square$

## 5 Dominated majorization

We have already pointed out that inequalities (1)-(2) were proven first for operator convex or operator concave functions, being extended only afterwards for ordinary convex/concave functions. Moreover, the proofs for ordinary convex/concave functions actually exploited the corresponding results for operator convex/concave functions. This may seem somewhat unnatural and it is not unreasonable to ask for a more direct proof.

In this section we introduce a number of new ideas and techniques which, although they may seem strange and somewhat contrived at first, will lead

to new, elementary proofs of inequalities (1)-(2) that bypass the Ando-Zhan theorem and do not require the machinery of operator monotone and operator convex functions. Secondly, we will use this technique to try and answer Question 2 raised in the previous section.

Let us consider three Hermitian matrices  $A$ ,  $B$  and  $C$  and assume that there exists  $a_0 > 0$  such that the following relation holds for all  $a \geq a_0$ , and for certain (possibly all) values of  $k$ :

$$\sum_{j=1}^k \lambda_j^\downarrow(aA + B) \leq \sum_{j=1}^k \lambda_j^\downarrow(aA + C). \quad (10)$$

As it holds for all  $a \geq a_0$ , it should be possible to simplify this condition.

Subtracting  $\sum_{j=1}^k \lambda_j^\downarrow(aA)$  from both sides, and substituting  $a = 1/t$ , we obtain

$$\frac{1}{t} \sum_{j=1}^k (\lambda_j^\downarrow(A + tB) - \lambda_j^\downarrow(A)) \leq \frac{1}{t} \sum_{j=1}^k (\lambda_j^\downarrow(A + tC) - \lambda_j^\downarrow(A)),$$

for all  $0 < t \leq t_0 = 1/a_0$ . In the limit of positive  $t$  going to 0, this yields a comparison between directional derivatives of sums of  $k$  largest eigenvalues:

$$\left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(A + tB) \leq \left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(A + tC). \quad (11)$$

Let us introduce the vector  $\delta(B; A)$  defined as:

$$\sum_{j=1}^k \delta_j(B; A) := \left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(A + tB). \quad (12)$$

With this notation, relation (11) becomes

$$\sum_{j=1}^k \delta_j(B; A) \leq \sum_{j=1}^k \delta_j(C; A).$$

That is, the entries of  $\delta(B; A)$  are related via a majorization-like relation (without the usual rearrangement) to those of  $\delta(C; A)$ .

To simplify the notations, we will use the symbol  $\prec_w$  for this relation:

$$a \prec_w b \iff \sum_{j=1}^k a_j \leq \sum_{j=1}^k b_j, \quad (13)$$

and explicitly put rearrangements in the vectors concerned by use of the symbols  $\uparrow$  and  $\downarrow$ . In that way, we write the classical majorization relation as  $a^\downarrow \prec_w b^\downarrow$ .

With these notations relation (11) is expressed as

$$\delta(B; A) \prec_w \delta(C; A). \quad (14)$$

We call this relation *A-dominated majorization* or *A-majorization* for short.

**Definition 1** Consider three Hermitian matrices  $A$ ,  $B$  and  $C$ . When the relation (11) holds, or equivalently, (14), we say that  $B$  is *A-majorized* by  $C$ .

The argument shown above proves the following:

**Proposition 1** Let  $A$ ,  $B$  and  $C$  be Hermitian matrices. If there exists  $a_0 > 0$  such that  $\sum_{j=1}^k \lambda_j^\downarrow(aA + B) \leq \sum_{j=1}^k \lambda_j^\downarrow(aA + C)$  holds for all  $a \geq a_0$ , then  $\delta(B; A) \prec_w \delta(C; A)$ .

### 5.1 Directional derivative of the sum of the $k$ -th largest eigenvalues

It turns out that there is a very simple way to calculate  $\delta(B; A)$ , based on an explicit expression of the directional derivative of the sum of the  $k$  largest eigenvalues of a symmetric matrix, which is well-known in numerical analysis (see [13] and references therein, and [20]). The directional derivative of a convex function is defined as follows ([13], Section 2.2):

**Definition 2** Let  $f(x)$  be a convex function defined on a subset  $\mathcal{O}$  of a Euclidean space  $X$ . For any  $x \in \mathcal{O}$ , and  $d \in X$ , the directional derivative of  $f$  at  $x$  in the direction  $d$  is defined as

$$f'(x, d) = \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t}.$$

It is essential that the limit  $t \rightarrow 0^+$  is taken because  $f$  need not be differentiable. We will denote this directional derivative by the symbol  $\left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+}$ .

Consider an  $n \times n$  Hermitian matrix  $A$ , and let its eigenvalues, sorted in non-increasing order, be denoted by  $\lambda_j^\downarrow(A)$ ,  $j = 1, 2, \dots, n$ . Let its *distinct* eigenvalues, sorted in decreasing order, be denoted by  $\mu_i(A)$ ,  $i = 1, 2, \dots, m$  (with  $m$  the number of distinct eigenvalues) and the corresponding multiplicities by  $r_i$ . Thus  $\sum_{i=1}^m r_i = n$ . The sum of the  $k$  largest eigenvalues of  $A$  can be written in terms of the  $\mu_i$  as follows: writing  $k$  as  $k = r_1 + r_2 + \dots + r_l + s$ ,

where  $1 \leq s \leq r_{l+1}$ ,

$$\sum_{j=1}^k \lambda_j^\downarrow(A) = \sum_{i=1}^l r_i \mu_i(A) + s \mu_{l+1}(A).$$

Furthermore, let  $P_i$  denote the projector onto the  $i$ -th eigenspace of  $A$ , corresponding to eigenvalue  $\mu_i(A)$ . Thus,  $P_i$  is a matrix of dimensions  $r_i \times n$ . The spectral decomposition of  $A$  can then be written as

$$A = \sum_{i=1}^m \mu_i(A) P_i^* P_i.$$

The following is a reformulation of Corollary 3.9 in [13], which was proven there for real symmetric matrices.

**Proposition 2** *Let  $A$  be a real  $n \times n$  symmetric matrix with spectral decomposition  $A = \sum_{i=1}^m \mu_i(A) P_i^* P_i$  and multiplicities  $r_i$ . Let  $B$  also be a real  $n \times n$  symmetric matrix. With  $k$  written as  $k = r_1 + r_2 + \dots + r_l + s$ , where  $1 \leq s \leq r_{l+1}$ , the directional derivative of  $\sum_{j=1}^k \lambda_j^\downarrow(A)$  in direction  $B$  is given by*

$$\left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(A + tB) = \sum_{i=1}^l \text{Tr } P_i B P_i^* + \sum_{j=1}^s \lambda_j^\downarrow(P_{l+1} B P_{l+1}^*). \quad (15)$$

Note that, when  $s = r_{l+1}$ , this formula simplifies to

$$\left. \frac{\partial}{\partial t} \right|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(A + tB) = \sum_{i=1}^{l+1} \text{Tr } P_i B P_i^*. \quad (16)$$

We summarise what we really need to know about this proposition in the following theorem (quietly extended to the complex case).

**Theorem 2** *Let  $A$  and  $B$  be Hermitian matrices. With  $\delta(B; A)$  defined by (12), the entries of the vector  $\delta(B; A)$  are the diagonal entries of  $B$  in a certain basis in which  $A$  is diagonal and its diagonal entries appear sorted in non-increasing order. When all eigenvalues of  $A$  are simple (i.e. have multiplicity 1), this basis is just the eigenbasis of  $A$  and does not depend on  $B$ .*

An independent proof of this theorem, that also works for complex Hermitian matrices, is presented in Section 7.

The upshot of Theorem 2 is that there exists a unitary matrix  $U$  such that  $U^* A U = \Lambda^\downarrow(A)$  and  $\delta(B; A) = \text{Diag}(U^* B U)$ . In other words,  $\delta(B; A)$  is the vector of diagonal elements of  $B$ , in a particular basis governed by  $A$ , and

possibly by  $B$  too. In the generic case that all  $\lambda_i(A)$  are distinct,  $U$  is unique and does not depend on  $B$ , hence in that case  $\delta(B; A)$  is the vector of diagonal elements of  $B$  in the eigenbasis of  $A$ .

## 5.2 Dominated majorization for co-diagonal matrices

Let us now specialise to the case where  $A$  and  $B$  commute and there is a common basis in which the diagonal elements of  $A$  and  $B$  appear in the same, non-increasing order. We will say that  $A$  and  $B$  that satisfy this condition are *co-diagonal*.

According to Proposition 1, validity of (10) for all  $a > 0$  implies  $A$ -majorization, (14). Theorem 2 now immediately leads to the following proposition, which says that for co-diagonal  $A$  and  $B$ , validity of (10) for all  $a > 0$  is actually equivalent with  $A$ -majorization.

**Proposition 3** *For Hermitian  $A, B, C$ , where  $A$  and  $B$  are co-diagonal, the following are equivalent:*

$$\lambda^\downarrow(aA + B) \prec_w \lambda^\downarrow(aA + C), \quad \forall a \geq 0 \quad (17)$$

$$\delta(B; A) \prec_w \delta(C; A) \quad (18)$$

$$\delta(aA + B; A) \prec_w \delta(aA + C; A), \quad \forall a \geq 0. \quad (19)$$

*Proof.*

(17) implies (18):

If relation (10) holds for all  $a > 0$ , then it holds for  $a$  tending to infinity. By Proposition 1 we then get that  $B$  is  $A$ -majorized by  $C$ .

(18) implies (19):

Let us add  $a\lambda^\downarrow(A)$  to both sides of (18). By Theorem 2,  $\delta(B; A)$  is the vector of diagonal elements of  $B$ , in a basis in which  $A$  is diagonal and the eigenvalues of  $A$  appear sorted in non-increasing order. Thus,  $\forall a > 0$ ,  $\delta(B; A) + a\lambda^\downarrow(A) = \delta(B + aA; A)$ . The same holds for  $C$ .

(19) implies (17):

By the co-diagonality of  $A$  and  $B$ ,  $aA + B$  is diagonal in any basis in which  $A$  is diagonal. Hence, the LHS of (19) is equal to  $\lambda^\downarrow(aA + B)$ . By Schur's majorization theorem, the RHS of (19) is majorized by  $\lambda^\downarrow(aA + C)$ .  $\square$

## 6 Applications of dominated majorization

In this section we first use Proposition 3, to give a new, elementary proof of inequality (1) for non-negative concave functions (which readily implies validity of inequality (2) for non-negative convex functions), that does not rely on the Ando-Zhan inequality for operator concave functions, nor on the theory of operator monotone functions.

Then, we answer Question 2 in the negative by exhibiting a counterexample. Here, too, Proposition 3 was instrumental.

### 6.1 A new proof of inequality (1) for non-negative concave functions

We want to prove that

$$|||f(A+B)||| \leq |||f(A) + f(B)|||$$

holds for all non-negative concave functions  $f(x)$ . Therefore, it should hold in particular for all functions  $f(x) = b + ax + f_0(x)$ , where  $f_0$  is non-negative concave with  $f_0(0) = 0$  and  $f'_0(+\infty) = 0$ , and for all  $a, b \geq 0$ . Inserting this in the eigenvalue-majorization form of inequality (1), we get the majorization relation

$$\lambda^\downarrow(b + a(A+B) + f_0(A+B)) \prec_w \lambda^\downarrow(2b + a(A+B) + f_0(A) + f_0(B)),$$

for  $A, B \geq 0$ . Clearly, this is strongest for  $b = 0$ . Proposition 3 then immediately yields the equivalent form

$$\delta(f(A+B); A+B) \prec_w \delta(f(A) + f(B); A+B),$$

for all non-negative concave functions  $f$  (recall that such functions are non-decreasing) with  $f(0) = 0$ .

An interesting aspect of this form is that, unlike  $\lambda$ ,  $\delta$  is linear in its first argument. Our proof of the equivalent form, stated as Proposition 4 below, crucially depends on this property.

**Proposition 4** *For positive semidefinite  $A$  and  $B$ , and  $f$  a non-negative concave function with  $f(0) = 0$ ,*

$$\delta(f(A+B); A+B) \prec_w \delta(f(A) + f(B); A+B). \quad (20)$$

*Proof.* Any non-negative concave function  $f$  can be uniformly approximated as a positive linear combination of angle functions  $x \mapsto x - (x-t)_+$ . By linearity

of  $\delta$ , inequality (20) follows if it holds for any such angle function, i.e.

$$\delta(A + B - (A + B - t)_+; A + B) \prec_w \delta(A - (A - t)_+ + B - (B - t)_+; A + B),$$

which, again by linearity, simplifies to

$$\delta((A - t)_+ + (B - t)_+; A + B) \prec_w \delta((A + B - t)_+; A + B).$$

In fact, for angle functions the latter inequality even holds with rearrangement, and we shall prove

$$\delta^\downarrow((A - t)_+ + (B - t)_+; A + B) \prec_w \delta^\downarrow((A + B - t)_+; A + B),$$

for all  $t \geq 0$ . Letting  $\text{tr}(x)$  denote the sum  $\sum_{i=1}^n x_i$  of  $x = (x_1, \dots, x_n)$ , this relation can be expressed in a well-known way as

$$\text{tr}(\delta((A - t)_+ + (B - t)_+; A + B) - s)_+ \leq \text{tr}(\delta((A + B - t)_+; A + B) - s)_+,$$

for all  $s$  (and  $t \geq 0$ ). Since both vectors  $\delta$  are non-negative it suffices to consider the case  $s \geq 0$ . In the eigenbasis of  $A + B$ ,  $A + B$  itself is of course diagonal, hence the RHS simplifies to  $\text{Tr}(A + B - (s + t))_+$ .

Now we introduce the variable  $u = s + t$ . The last inequality has to be valid for all values of  $s$  and  $t$ , thus if we keep the value of  $u$  fixed, the inequality has to remain true if we maximise the LHS over all values of  $t$  in the range  $[0, u]$  (and set  $s = u - t$ ). That is,

$$\begin{aligned} & \max_{0 \leq t \leq u} \text{tr}(\delta((A - t)_+ + (B - t)_+; A + B) - u + t)_+ \\ & \leq \text{Tr}(A + B - u)_+. \end{aligned} \tag{21}$$

The next important consequence of the simple behaviour of  $\delta$  is that the function  $t \mapsto F(t) := \text{tr}(\delta((A - t)_+ + (B - t)_+; A + B) - u + t)_+$  is convex. Note first that the positive part function is convex and increasing. Applying this to its outer appearance in the definition of  $F$ , the required convexity of  $F(t)$  follows if, for any  $i$ ,  $\delta((A - t)_+ + (B - t)_+; A + B)_i - u + t$  is itself a convex function of  $t$ . This function can be written as  $((A - t)_+)_{ii} + ((B - t)_+)_{ii} - u + t$ , in the eigenbasis of  $A + B$ . Hence, convexity follows from the convexity of  $t \mapsto (\psi, (A - t)_+ \psi)$ , for any vector  $\psi$ , and to see the latter, just consider this quantity in the eigenbasis of  $A$  and see that it can be written as  $\sum_{j=1}^n (\lambda_j(A) - t)_+ |\psi_j|^2$ , which is a positive linear combination of angle functions and, therefore, convex.

The convexity of  $F(t)$  now implies the simple fact that the maximum in the LHS of (21)  $\max_{0 \leq t \leq u} \text{tr}(\delta((A - t)_+ + (B - t)_+; A + B) - u + t)_+$  is achieved in one of the extreme points, either in  $t = 0$  or in  $t = u$ . Noting that  $A$  and  $B$  are



positive semidefinite, the value achieved in  $t = 0$  is  $\text{tr}(\delta(A + B; A + B) - u)_+$ , which is identical to the RHS in (21). It therefore only remains to show that the value in  $t = u$  is also bounded above by the RHS. Using the fact that the function  $\text{tr} \delta(X; Y)$  is always equal to  $\text{Tr} X$ , this amounts to the inequality

$$\text{Tr}(A - u)_+ + \text{Tr}(B - u)_+ \leq \text{Tr}(A + B - u)_+. \quad (22)$$

Here, the outer appearance of the positive part function in the LHS has been removed because its argument is always positive semidefinite.

To prove inequality (22), recall the norm inequality

$$|||A \oplus B||| \leq |||(|A| + |B|) \oplus 0|||,$$

valid for any unitarily invariant norm ([7], Theorem IV.2.13). In particular, it holds for the Ky Fan norms, and for PSD  $A$  and  $B$  can be written as the eigenvalue majorization

$$\lambda^\downarrow(A \oplus B) \prec_w \lambda^\downarrow((A + B) \oplus 0).$$

Thus, for all  $u \geq 0$  (again, by non-negativity of  $A$  and  $B$ , it suffices to consider  $u \geq 0$ ),

$$\text{Tr} \left( \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} - u \right)_+ \leq \text{Tr} \left( \begin{pmatrix} A + B & 0 \\ 0 & 0 \end{pmatrix} - u \right)_+,$$

which is nothing but inequality (22), reformulated in terms of  $2 \times 2$  block matrices. This ends the proof of the proposition.  $\square$

One might still object that our proof is not really elementary, relying as it is on Proposition 3 and the theory behind it. Strictly speaking, though, Proposition 3 is not needed in the proof, and only provided the intuition to try and prove the equivalent form (20). Indeed, validity of inequality (1) follows immediately from Proposition 4 by combining it with Schur's majorization theorem:

$$\begin{aligned} \lambda^\downarrow(f(A + B)) &= \delta(f(A + B); A + B) \\ &\prec_w \delta(f(A) + f(B); A + B) \\ &\prec_w \lambda^\downarrow(f(A) + f(B)). \end{aligned}$$

As already shown by Ando and Zhan [2], validity of inequality (1) for a given non-negative increasing concave function  $f$  implies inequality (2) for the inverse function  $g = f^{-1}$ . Hence, in combination with our proof of inequality (1), this also yields an elementary proof of inequality (2) for non-negative convex functions  $g(x)$  with  $g(0) = 0$ . This was first proven independently from (1)

by Kosem, by appealing to the corresponding inequality for operator convex functions.

For completeness, we repeat the short Ando-Zhan argument here.

*Proof of inequality (2) for non-negative convex functions.* Let  $g(x)$  be a non-negative convex function with  $g(0) = 0$ . Thus  $g$  is increasing. In particular,  $g$  applied to vectors is strongly isotone [7].

Let  $f(x)$  be its inverse function,  $f = g^{-1}$ ; thus  $f(x)$  is a non-negative increasing concave function with  $f(0) = 0$ . For such  $f$ , we have (inequality (1))

$$\lambda^\downarrow(f(A + B)) \prec_w \lambda^\downarrow(f(A) + f(B)).$$

Since  $g(x)$  is strongly isotone, applying  $g$  on both sides preserves weak majorization:

$$g(\lambda^\downarrow(f(A + B))) \prec_w g(\lambda^\downarrow(f(A) + f(B))).$$

This simplifies, by monotonicity of  $g$ , to

$$\lambda^\downarrow(A + B) \prec_w \lambda^\downarrow(g(f(A) + f(B))).$$

Substituting  $g(A)$  for  $A$  and  $g(B)$  for  $B$  then yields inequality (2).  $\square$

## 6.2 Counterexample to Question 2

To answer Question 2, we will first disregard the absolute values and consider the property that a convex function  $f$  satisfies

$$\lambda(f(\Delta)) \prec_w \lambda(f(B + \Delta) - f(B)) \tag{23}$$

for all PSD  $B$  and  $\Delta$ , which is equivalent to the statement

$$\lambda(f(A - B)) \prec_w \lambda(f(A) - f(B)) \tag{24}$$

for all  $A \geq B \geq 0$ .

Although it is by no means obvious at this point, when  $\Delta > 0$  strictly, Question 2 is equivalent to validity of (23) for all stated functions. While it is obvious that (23) implies  $|||g(B + \Delta) - g(B)||| \geq |||g(\Delta)|||$ , the opposite is not necessarily true because of the absolute value implicit in the definition of the norm. Nevertheless, it will turn out that a counterexample to (23) for some function  $g$  will indirectly yield a counterexample to Question 2 for *some other function*  $\tilde{g}(x) = g(x) + \alpha x$ , with  $\alpha > 0$  large enough, provided  $\Delta > 0$  holds strictly.

Here,  $\alpha$  must be large enough to make  $\tilde{g}(B+\Delta) - \tilde{g}(B) = g(B+\Delta) - g(B) + \alpha\Delta$  positive semidefinite, in which case the absolute value signs can be left out. This will all be made clear below.

The monotone convex angle functions  $x \mapsto ax + (x - 1)_+$  ( $a \geq 0$ ) already have proven their valour as a testing ground for similar statements, in Section 3. Numerical experiments using angle functions for inequality (23) did not directly lead to any counterexamples, however. This temporarily increased our belief that the inequality might actually hold, and led us to investigate, as an initial step towards a ‘proof’, whether the inequality

$$\sum_{j=1}^k \lambda_j^\downarrow(aY + B) \leq \sum_{j=1}^k \lambda_j^\downarrow(aY + C)$$

might be true for all  $a \geq 0$ , where  $B = f(Y)$  and  $C = f(X + Y) - f(X)$ , and  $f(x) = (x - 1)_+$ .

If the answer to Question 2 is to be affirmative, it should at least hold for all angle functions  $f(x) = ax + b(x - x_0)_+$ . By Proposition 3 this is equivalent to the statement

$$\delta((Y - \mathbf{I})_+; Y) \prec_w \delta((X + Y - \mathbf{I})_+ - (X - \mathbf{I})_+; Y).$$

Consider the  $3 \times 3$  PSD matrices

$$X = \begin{pmatrix} 0.35614 & -0.053243 & 0.10116 \\ -0.053243 & 0.87456 & 0.40559 \\ 0.10116 & 0.40559 & 0.82474 \end{pmatrix}$$

and

$$Y = \begin{pmatrix} 0.53642 & 0 & 0 \\ 0 & 0.42018 & 0 \\ 0 & 0 & 0.094866 \end{pmatrix}.$$

The eigenbasis of  $Y$  is therefore the standard basis. Then  $\delta((Y - \mathbf{I})_+; Y) = (0, 0, 0)$  and

$$(X + Y - \mathbf{I})_+ - (X - \mathbf{I})_+ = \begin{pmatrix} -0.00018194 & 0.00052449 & -0.0016345 \\ 0.00052449 & 0.2573 & 0.12368 \\ -0.0016345 & 0.12368 & 0.04 \end{pmatrix}$$

so that  $\delta((X + Y - \mathbf{I})_+ - (X - \mathbf{I})_+; Y) = (-0.00018194, 0.2573, 0.04)$ . The first entry is negative, violating the majorization relation.

Now, as mentioned above, this counterexample immediately yields a counterexample to Question 2. Consider thereto the function  $f(x) = \alpha x + (x - 1)_+$  with  $\alpha = 1$ , say. Then the LHS of the inequality becomes  $\delta(Y + (Y - \mathbf{I})_+; Y) = (0.53642, 0.42018, 0.094866)$  and the RHS  $\delta(Y + (X + Y - \mathbf{I})_+ - (X - \mathbf{I})_+; Y) = (0.53624, 0.67748, 0.13487)$ , again violating the inequality. Since  $Y + (X + Y - \mathbf{I})_+ - (X - \mathbf{I})_+$  is a positive definite matrix (as can be checked numerically), it is unchanged by putting in the required absolute value signs.

Even more explicitly, consider the function  $g(x) = 101x + (x - 1)_+$ . Then  $\lambda^\downarrow(g(X + Y) - g(X)) = (54.17824, 42.69595, 9.621004)$  while  $\lambda^\downarrow(g(Y)) = (54.17842, 42.43818, 9.581466)$ . This clearly violates the eigenvalue majorization relation of Question 2, *with* absolute value signs, because of the positivity of  $g(X + Y) - g(X)$ .

## 7 Proof of Theorem 2

In this section, we give a self-contained proof of Theorem 2 that does not rely on the methods of convex analysis and is also valid for complex Hermitian matrices, not only real-symmetric ones. For convenience, we reformulate the statement of the theorem here.

Define a *proper eigenbasis* of a Hermitian matrix  $A$  as an orthonormal basis in which  $A$  is diagonal and its diagonal entries are the eigenvalues of  $A$  sorted in non-increasing order.

**Theorem 2'.** Let  $A$  and  $B$  be Hermitian matrices. With  $\delta(B; A)$  defined via equation (12), the entries of the vector  $\delta(B; A)$  are the diagonal entries of  $B$  in some proper eigenbasis of  $A$ . When all eigenvalues of  $A$  are simple (i.e. have multiplicity 1), this proper eigenbasis is unique; otherwise the required one depends on  $B$ .

We need a number of definitions first, and recall some basic facts about the perturbation theory of eigenvalue decompositions (see, e.g. [14], Chapter 2, Section 1).

Consider the matrix-valued function  $z \mapsto A + zB$ ,  $z \in \mathbb{C}$ , with  $A$  and  $B$  the  $n \times n$  Hermitian matrices of the theorem. It is well-known that the roots of the characteristic function of  $A + zB$  are analytic functions of  $z$  with only algebraic singularities. This means that the number  $m$  of (distinct) eigenvalues of  $A + zB$  is a constant of  $z$ , with the exception of a number of special values of  $z$ , which will be called exceptional points. If  $m < n$ , we say that  $A + zB$  is permanently degenerate. In the exceptional points some of the eigenvalues may coincide; this is called an accidental degeneracy.

In the following we will consider a simply-connected subdomain  $D$  of the complex plane  $\mathbb{C}$  containing no exceptional points, and such that the intersection of  $D$  with the real axis is the interval  $(0, t_0)$ , with  $t_0 > 0$ . The closure of  $D$  is denoted  $\overline{D}$ , and its intersection with the real axis is  $[0, t_0]$ .

We can write the (possibly multiple) eigenvalues of  $A + zB$ ,  $z \in \overline{D}$ , as holomorphic functions  $\lambda_1(z), \lambda_2(z), \dots, \lambda_n(z)$ . For  $z = t \in \mathbb{R}$ , these eigenvalues are real and can be sorted. Sorted in non-increasing order they will be denoted as  $\lambda_1^\downarrow(t) \geq \lambda_2^\downarrow(t) \geq \dots \geq \lambda_n^\downarrow(t)$ ,  $0 \leq t \leq t_0$ .

Furthermore, we can write the *distinct* eigenvalues of  $A + zB$ ,  $z \in D$ , as a fixed number of holomorphic functions  $\mu_1(z), \mu_2(z), \dots, \mu_m(z)$ . We will number them such that  $\mu_1(t) > \mu_2(t) > \dots > \mu_m(t)$  holds for  $t \in (0, t_0)$  (or  $t \in [0, t_0]$ ). We denote the multiplicity of  $\mu_i$  by  $r_i$ . Thus,  $n = \sum_{i=1}^m r_i$ .

The projector on the eigenspace of  $A + zB$  corresponding to  $\mu_i(z)$  will be denoted by the function  $\mathcal{P}_i(z)$ ,  $z \in D$ , and is called the eigenprojection for  $\mu_i(z)$ . This function is holomorphic on  $D$  [14].

If  $z = 0$  is not an exceptional point, the distinct eigenvalues of  $A$  are equal to the limiting values  $\mu_i(0)$ , and the corresponding eigenprojections coincide with the  $\mathcal{P}_i(0)$ .

If  $z = 0$  is an exceptional point then an accidental degeneracy occurs and  $A$  has less than  $m$  distinct eigenvalues. Each of these eigenvalues may split into several  $\mu_i(t)$ ; that is,  $\lim_{t \rightarrow 0} \mu_i(t) = \lambda$  for several (contiguous) values of  $i$ , say  $i = i_1, \dots, i_2$ , where  $\lambda$  is a certain eigenvalue of  $A$ . In that case, the eigenprojection for  $\lambda$  of  $A$  coincides with the sum  $\lim_{t \rightarrow 0} \sum_{j=i_1}^{i_2} \mathcal{P}_j(t)$ ; i.e. the  $\lim_{t \rightarrow 0} \mathcal{P}_j(t)$  separately are not themselves eigenprojectors of  $A$ .

Let  $k$  be an integer such that there exists an  $l$  for which  $k = r_1 + r_2 + \dots + r_l$ ; we shall say that such a  $k$  is an *entire sum* of the multiplicities  $r_i$ . For such values of  $k$ , we define the projector  $\mathcal{P}_{(k)}(z)$  as the sum of eigenprojectors

$$\mathcal{P}_{(k)}(z) = \mathcal{P}_1(z) + \mathcal{P}_2(z) + \dots + \mathcal{P}_l(z).$$

For  $z = t$  real, this is the projector on the subspace spanned by the eigenvectors of the  $k$  largest eigenvalues (counting multiplicities) of  $A + tB$ . Since the  $\mathcal{P}_i(z)$  are holomorphic functions on  $D$ , so is  $\mathcal{P}_{(k)}(z)$ . By continuity of the eigenvalues  $\lambda_k(z)$ , we have for any such  $k$ ,

$$\sum_{j=0}^k \lambda_k^\downarrow(A) = \lim_{t \rightarrow 0^+} \sum_{j=0}^k \lambda_k^\downarrow(t) = \text{Tr}[\lim_{t \rightarrow 0^+} \mathcal{P}_{(k)}(t) A].$$

If  $k$  cannot be written in this way, i.e.  $k = r_1 + r_2 + \dots + r_l + s$  with  $s$

a ‘remainder’ satisfying  $0 < s < r_{l+1}$ , we cannot uniquely define  $\mathcal{P}_{(k)}(z)$ , because there is an infinity of  $s$ -dimensional subspaces in the eigenspace for  $\mu_{l+1}$ . Hence, we will only define  $\mathcal{P}_{(k)}(z)$  for  $k$  that are entire sums of  $r_i$ .

Finally, we define the projectors  $\mathcal{P}_{(k)}$ . If  $z = 0$  is not an exceptional point, and  $k$  is an entire sum of multiplicities  $r_i$ , then  $\mathcal{P}_{(k)}(0)$  is defined, and we define  $\mathcal{P}_{(k)} := \mathcal{P}_{(k)}(0)$ . If  $z = 0$  is an exceptional point then  $A + zB$  has an accidental degeneracy at  $z = 0$ . Even if  $k$  is an entire sum of multiplicities  $r_i$  of  $A + zB$ , it need not be an entire sum of multiplicities of  $A$ . Hence, in that case  $\mathcal{P}_{(k)}(t)$  is only defined for  $t \in (0, t_0)$  (with  $t_0 > 0$ ) but not for  $t = 0$ . We will then define  $\mathcal{P}_{(k)}$  as the limiting value

$$\mathcal{P}_{(k)} := \lim_{t \rightarrow 0^+} \mathcal{P}_{(k)}(t).$$

For all other values of  $k$ ,  $\mathcal{P}_{(k)}$  will not be defined.

**Lemma 1** *If  $k$  is such that  $\mathcal{P}_{(k)}$  is defined (directly in  $t = 0$  or via the limit  $t \rightarrow 0^+$ ), then*

$$\sum_{j=1}^k \delta_j(B; A) = \text{Tr } B \mathcal{P}_{(k)}.$$

*Proof.* Consider the variational characterization of the sum of the  $k$  largest eigenvalues of a Hermitian matrix  $H$ :

$$\sum_{j=1}^k \lambda_j^\downarrow(H) = \max_Q \text{Tr}[H Q],$$

where  $Q$  runs over all rank- $k$  projectors. If  $k$  is such that  $\mathcal{P}_{(k)}(H)$  exists (taking the potential degeneracies of  $H$  into account) then  $Q = \mathcal{P}_{(k)}(H)$  achieves the maximum, i.e.  $\max_Q \text{Tr}[H Q] = \text{Tr}[H \mathcal{P}_{(k)}(H)]$ .

We have, in particular, that  $\mathcal{P}_{(k)}(t) := \mathcal{P}_{(k)}(A + tB)$  (if it exists) achieves the maximum for  $H = A + tB$ . More precisely, for any  $t$  in the open interval  $(0, t_0)$ , the function  $u \mapsto \text{Tr}[(A + tB) \mathcal{P}_{(k)}(u)]$  achieves its maximum over  $(0, t_0)$  in the *interior* point  $u = t$ . Since  $\mathcal{P}_{(k)}(t)$  is holomorphic, this function is differentiable, hence this maximum must be a stationary point. Thus

$$\left. \frac{\partial}{\partial u} \right|_{u \rightarrow t} \text{Tr}[(A + tB) \mathcal{P}_{(k)}(u)] = 0,$$

i.e.

$$\text{Tr}[(A + tB) \frac{\partial}{\partial t} \mathcal{P}_{(k)}(t)] = 0.$$

This implies

$$\begin{aligned}
\frac{\partial}{\partial t} \sum_{j=1}^k \lambda_j^\downarrow(t) &= \frac{\partial}{\partial t} \text{Tr}[(A + tB) \mathcal{P}_{(k)}(t)] \\
&= \text{Tr}[(A + tB) \frac{\partial}{\partial t} \mathcal{P}_{(k)}(t)] + \text{Tr}[B \mathcal{P}_{(k)}(t)] \\
&= \text{Tr}[B \mathcal{P}_{(k)}(t)].
\end{aligned}$$

In particular,

$$\sum_{j=1}^k \delta_j(A; B) = \frac{\partial}{\partial t} \Big|_{t \rightarrow 0^+} \sum_{j=1}^k \lambda_j^\downarrow(t) = \lim_{t \rightarrow 0^+} \text{Tr}[B \mathcal{P}_{(k)}(t)] = \text{Tr}[B \mathcal{P}_{(k)}].$$

□

We are now in the position to prove Theorem 2. Let's first consider the simplest case when  $A$  is not degenerate, i.e. all eigenvalues of  $A + zB$  are simple for  $z \in \overline{D}$ . In that case  $\mathcal{P}_{(k)}(z)$  is always defined for all  $k$  and all  $z \in \overline{D}$ , and, hence,  $\mathcal{P}_{(k)}$  is defined as  $\mathcal{P}_{(k)}(0) = \sum_{j=1}^k \mathcal{P}_j(0)$ . There is a unique unitary matrix  $U$  such that  $UAU^* = \text{Diag}(\lambda^\downarrow(A))$ , and in this basis the projector  $\mathcal{P}_j$  is expressed as  $e^{jj}$ . Hence, by the lemma we have that  $\sum_{j=1}^k \delta_j(B; A) = \text{Tr} B \mathcal{P}_{(k)} = \sum_{j=1}^k B_{jj}$ , where the  $B_{jj}$  are the diagonal elements of  $B$  expressed in that same basis. Therefore, for all  $j$ ,  $\delta_j(B; A) = B_{jj}$ .

If  $A$  is degenerate, there is no unique eigenbasis of  $A$ . However, the lemma only requires us to deal with the limits  $\lim_{t \rightarrow 0^+} \mathcal{P}_{(k)}(t)$ . If the degeneracy of  $A$  is lifted completely in  $A + zB$ , i.e. all eigenvalues of  $A$  split into simple eigenvalues, then all  $\mathcal{P}_j(t)$  are rank-1 projectors and  $\mathcal{P}_{(k)}(t) = \sum_{j=1}^k \mathcal{P}_j(t)$ . Furthermore, letting  $i_1$  and  $i_2$  be any pair of indices such that an eigenvalue of  $A$  splits into the eigenvalues  $\mu_{i_1}, \dots, \mu_{i_2}$  of  $A + zB$ , we have that  $\lim_{t \rightarrow 0} \sum_{j=i_1}^{i_2} \mathcal{P}_j(t)$  is an eigenprojector of  $A$ . Therefore, there exists a unique proper eigenbasis of  $A$  (determined by  $B$ ) in which  $\lim_{t \rightarrow 0} \mathcal{P}_j(t) = e^{jj}$ , the elementary matrix with a 1 in position  $(j, j)$  and zeroes elsewhere. Again we find that, for all  $j$ ,  $\delta_j(B; A) = B_{jj}$  in that proper eigenbasis.

The most complicated case arises when  $A + zB$  is permanently degenerate, i.e. the degeneracies are not lifted completely, as some eigenvalues of  $A$  may split into still degenerate eigenvalues  $\mu_i$  of  $A + zB$ , with multiplicities  $r_i$ . Then the projectors  $\mathcal{P}_i(z)$  have rank  $r_i$ , and the  $\mathcal{P}_{(k)}(t)$  are only defined when  $k$  is an entire sum of the multiplicities  $r_i$ . There still exists a proper eigenbasis of  $A$  in which the projectors  $\lim_{t \rightarrow 0} \mathcal{P}_j(t)$  are diagonal, now of the form  $0 \oplus \mathbf{I}_{r_j} \oplus 0$ , but it is no longer unique; we will exploit exactly this freedom to deal with  $k$  that are not entire sums.

If  $k$  is not an entire sum of  $r_i$ , we have  $k = r_1 + r_2 + \dots + r_l + s$ , with  $s$  the remainder term, satisfying  $1 \leq s < r_{l+1}$ . We first write  $k$  as an interpolated value between two entire sums as follows:

$$\begin{aligned}
k &= \frac{s}{r_{l+1}}(r_1 + r_2 + \dots + r_{l+1}) + (1 - \frac{s}{r_{l+1}})(r_1 + r_2 + \dots + r_l) \\
&= \alpha k_+ + (1 - \alpha)k_-.
\end{aligned}$$

Here we defined  $\alpha = s/r_{l+1}$ , and the two entire sums  $k_- = r_1 + r_2 + \dots + r_l$  and  $k_+ = r_1 + r_2 + \dots + r_{l+1}$ . We can express  $\sum_{j=1}^k \lambda_j^\downarrow$  as a linear interpolation between  $\sum_{j=1}^{k_-} \lambda_j^\downarrow$  and  $\sum_{j=1}^{k_+} \lambda_j^\downarrow$ :

$$\begin{aligned}
\sum_{j=1}^k \lambda_j^\downarrow(t) &= \sum_{i=1}^l r_i \mu_i(t) + s \mu_{l+1}(t) \\
&= \text{Tr}[(A + tB) (\mathcal{P}_1(t) + \dots + \mathcal{P}_l(t) + \frac{s}{r_{l+1}} \mathcal{P}_{l+1}(t))] \\
&= \text{Tr}[(A + tB) (\alpha \mathcal{P}_{(k_+)}(t) + (1 - \alpha) \mathcal{P}_{(k_-)}(t))] \\
&= \alpha \text{Tr}[(A + tB) \mathcal{P}_{(k_+)}(t)] + (1 - \alpha) \text{Tr}[(A + tB) \mathcal{P}_{(k_-)}(t)].
\end{aligned}$$

Applying the Lemma to both terms, we obtain

$$\sum_{j=1}^k \delta_j(B; A) = \text{Tr}[B(\alpha \mathcal{P}_{(k_+)} + (1 - \alpha) \mathcal{P}_{(k_-)})] = \sum_{i=1}^l \text{Tr} B \mathcal{P}_i + \alpha \text{Tr} B \mathcal{P}_{l+1}. \quad (25)$$

Again, to deal with eigenvalue splitting at  $z = 0$ , each of the  $\mathcal{P}_i$  corresponds to the limit  $\lim_{t \rightarrow 0^+} \mathcal{P}_i(t)$ .

Let us consider a partitioning of  $B$  in an eigenbasis of  $A + zB$  mentioned before, in which the  $\mathcal{P}_i(z)$  appear in the form  $0 \oplus \mathbf{I}_{r_i} \oplus 0$ . That is, in  $B$  we can single out blocks on its diagonal, each of which corresponds to an eigenspace of  $A + zB$ ; Then  $\text{Tr} B \mathcal{P}_i(z)$  is the sum of all  $r_i$  diagonal elements of the  $i$ -th block of  $B$ .

The degeneracy of the eigenvalues  $\mu_i(z)$  means that this eigenbasis is still not unique and is determined up to ‘local’ rotations within each of the eigenspaces. We can use this freedom to make the diagonal elements of  $B$  equal within each block. This allows us to get rid of  $\alpha$  in (25). Indeed, as  $\alpha = s/r_{l+1}$  and  $\text{Tr} B \mathcal{P}_{l+1}$  is the sum of all  $r_{l+1}$  diagonal elements of the  $(l + 1)$ -th block of  $B$ , then if all these diagonal elements are equal,  $\alpha \text{Tr} B \mathcal{P}_{l+1}(z)$  is equal to the sum of the first  $s$  diagonal elements of  $B$  in that block.

Wrapping up we find that  $\sum_{i=1}^l \text{Tr} B \mathcal{P}_i + \alpha \text{Tr} B \mathcal{P}_{l+1}$  equals the sum of the first  $r_1 + r_2 + \dots + r_l + s = k$  diagonal elements of  $B$  in the chosen eigenbasis. Taking the limit  $z = t \rightarrow 0$ , we finally obtain that, again, there is a proper eigenbasis of  $A$  in which

$$\sum_{j=1}^k \delta_j(B; A) = \sum_{j=1}^k B_{jj},$$



and hence  $\delta_j(B; A) = B_{jj}$ .  $\square$

KA acknowledges the hospitality of the Institut Mittag-Leffler, Djursholm (Sweden), where the final stages of the work have been done. We thank an anonymous referee for a variety of detailed comments, which helped to improve the exposition considerably.

## References

- [1] T. Ando, “Comparison of norms  $|||f(A) - f(B)|||$  and  $|||f(|A - B|)|||$ ,” *Math. Z.* **197**, 403–409 (1988).
- [2] T. Ando, X. Zhan, “Norm inequalities related to operator monotone functions,” *Math. Ann.* **315**, 771–780 (1999).
- [3] J. S. Aujla, “Some norm inequalities for completely monotone functions,” *Siam J. Matrix Anal. Appl.* **22**, 569–573 (2000).
- [4] J. S. Aujla, J.-C. Bourin, “Eigenvalue inequalities for convex and log-convex functions,” *Linear Algebra Appl.* **424**, 25–35 (2007).
- [5] J. S. Aujla, F. C. Silva, “Weak majorization inequalities and convex functions,” *Linear Algebra Appl.* **369**, 217–233 (2003).
- [6] R. Bhatia, “Some inequalities for norm ideals,” *Comm. Math. Phys.* **111**, 33–39 (1987).
- [7] R. Bhatia, *Matrix Analysis*, Springer, Heidelberg (1997).
- [8] R. Bhatia, F. Kittaneh, “Norm inequalities for positive operators,” *Lett. Math. Phys.* **43**, 225–231 (1998).
- [9] M. Sh. Birman, L. S. Koplienko, M. Z. Solomyak, “Estimates of the spectrum of the difference between fractional powers of self-adjoint operators,” *Izvestiya Vysshikh Uchebnykh Zavedenni. Mat.* **19**, 3–10 (1975).
- [10] J.-C. Bourin, M. Uchiyama, “A matrix subadditivity inequality for  $f(A + B)$  and  $f(A) + f(B)$ ,” *Linear Algebra Appl.* **423**, 512–518 (2007).
- [11] F. Hiai, “Log-majorizations and norm inequalities for exponential operators,” *Banach Center Publ.* **38**, 119–181 (1997).
- [12] E. Hille and R.S. Phillips, *Functional analysis and semi-groups*, AMS Colloquium Publications vol. **31** (1957).
- [13] J.-B. Hiriart-Urruty and D. Ye, “Sensitivity analysis of all eigenvalues of a symmetric matrix,” *Numer. Math.* **70**, 45–72 (1995).

- [14] T. Kato, *Perturbation theory for linear operators*, Reprint of the 1980 edition, Classics in Mathematics, Springer-Verlag, Berlin (1995).
- [15] F. Kittaneh, “Inequalities for Schatten  $p$ -norm IV,” *Comm. Math. Phys.* **106**, 581–585 (1986).
- [16] F. Kittaneh, H. Kosaki, “Inequalities for Schatten  $p$ -norm V,” *Publ. Res. Inst. Math. Sci.* **23**, 433–443 (1987).
- [17] T. Kosem, “Inequalities between  $\|f(A + B)\|$  and  $\|f(A) + f(B)\|$ ,” *Linear Algebra Appl.* **418**, 153–160 (2006).
- [18] R. Mathias, “Concavity of monotone matrix functions of finite order,” *Linear and Multilinear Algebra* **27**, 129–138 (1990).
- [19] C. A. McCarthy, “ $c_p$ ,” *Israel J. Math.* **5**, 249–271 (1967).
- [20] M.L. Overton and R.S. Womersley, “Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices,” *Math. Programming* **62** Ser. B, 321–357 (1993).
- [21] R. T. Powers, E. Størmer, “Free states of the canonical anticommutator relations,” *Comm. Math. Phys.* **16**, 1–33 (1970).